Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **493,** 3178–3193 (2020)
Advance Access publication 2020 February 4

doi:10.1093/mnras/staa325

Deep transfer learning for star cluster classification: I. application to the PHANGS–*HST* survey

Wei Wei,[1,2★] E. A. Huerta,[1,3★] Bradley C. Whitmore,[4★] Janice C. Lee,[5] Stephen Hannon,[6] Rupali Chandar,[7] Daniel A. Dale,[8] Kirsten L. Larson,[5] David A. Thilker,[9] Leonardo Ubeda,[4] Médéric Boquien,[10] Mélanie Chevance,[11] J. M. Diederik Kruijssen,[11] Andreas Schruba,[12] Guillermo A. Blanc[13,14] and Enrico Congiu[13,15]

Affiliations are listed at the end of the paper

Accepted 2020 February 1. Received 2020 February 1; in original form 2019 August 29

ABSTRACT

We present the results of a proof-of-concept experiment that demonstrates that deep learning can successfully be used for production-scale classification of compact star clusters detected in *Hubble Space Telescope* (*HST*) ultraviolet-optical imaging of nearby spiral galaxies ($D \lesssim 20$ Mpc) in the Physics at High Angular Resolution in Nearby GalaxieS (PHANGS)–HST survey. Given the relatively small nature of existing, human-labelled star cluster samples, we transfer the knowledge of state-of-the-art neural network models for real-object recognition to classify star clusters candidates into four morphological classes. We perform a series of experiments to determine the dependence of classification performance on neural network architecture (ResNet18 and VGG19-BN), training data sets curated by either a single expert or three astronomers, and the size of the images used for training. We find that the overall classification accuracies are not significantly affected by these choices. The networks are used to classify star cluster candidates in the PHANGS–*HST* galaxy NGC 1559, which was not included in the training samples. The resulting prediction accuracies are 70 per cent, 40 per cent, 40–50 per cent, and 50–70 per cent for class 1, 2, 3 star clusters, and class 4 non-clusters, respectively. This performance is competitive with consistency achieved in previously published human and automated quantitative classification of star cluster candidate samples (70–80 per cent, 40–50 per cent, 40–50 per cent, and 60–70 per cent). The methods introduced herein lay the foundations to automate classification for star clusters at scale, and exhibit the need to prepare a standardized data set of human-labelled star cluster classifications, agreed upon by a full range of experts in the field, to further improve the performance of the networks introduced in this study.

Key words: galaxies: star clusters: general.

1 INTRODUCTION

Human visual classification of electromagnetic signals from astronomical sources is a core task in observational research with a long established history (Cannon & Pickering 1912, 1918; Hubble 1926, 1936; de Vaucouleurs 1963). It has been an essential means by which progress has been made in understanding the formation and evolution of structures from stars to galaxies. However, in the modern era of 'Big Data' in Astronomy, with unprecedented growth in electromagnetic survey area, field of view, sensitivity, resolution, wavelength coverage, cadence, and transient alert production, it has become apparent that human classification is no longer scalable (LSST Science Collaboration 2009; Abbott et al. 2016). This realization has motivated the use of machine learning techniques to automate image classification (Ball et al. 2008; Banerji et al. 2010; Carrasco Kind & Brunner 2013; Kamdar, Turk & Brunner 2016; Ishak 2017; Kim & Brunner 2017). Some of these machine learning algorithms have been integrated into widely used methods for image processing, such as the neural networks (NNs) trained for star/galaxy separation in the automated source detection and photometry software SEXTRACTOR (Bertin & Arnouts 1996). Other applications of machine learning for image classification include the

★ E-mail: weiw2@illinois.edu (WW); elihu@illinois.edu (EAH); whitmore@stsci.edu (BCW)

use of so-called decision trees (Weir, Fayyad & Djorgovski 1995; Suchkov, Hanisch & Margon 2005; Ball et al. 2006; Vasconcellos et al. 2011; Sevilla-Noarbe & Etayo-Sotos 2015) and support vector machines(Fadely, Hogg & Willman 2012; Małek et al 2013; Solarz et al. 2017).

Visual object recognition has also been a core research activity in the computer science community. For instance, the PASCAL VOC challenge was initiated to develop software to accurately classify about 20 000 images divided into 20 object classes (Everingham et al. 2015). Over the last decade, deep learning algorithms have rapidly evolved to become the state-of-the-art signal-processing tools for computer vision, to the point of surpassing human performance. The success of deep learning algorithms for image classification can be broadly attributed to the combination of increasing processing speed and the availability of very large data sets for training; i.e. graphics processing units (GPUs) to train, validate, and test NN models; and curation of high-quality, human-labelled data sets, such as the `ImageNet` data set (Deng et al. 2009), which has over 14 million images divided into more than 1000 object categories.

The `ImageNet` large scale visual recognition challenge (Russakovsky et al. 2015) has driven the development of deep learning models that have achieved breakthroughs for image classification. In 2012, the network architecture `AlexNet` (Krizhevsky, Sutskever & Hinton 2012) achieved a $\sim$ 50 per cent reduction in error rate in the `ImageNet` challenge – a remarkable feat at that time that relied on the use of GPUs for the training of the model, data augmentation (image translations, horizontal reflections, and mean subtraction), as well as other novel algorithm improvements that are at the core of state-of-the-art NN models today, e.g. using successive convolution and pooling layers followed by fully connected layers at the end of the NN architecture.

Within the next 2 yr, the architectures `VGGNet` (Simonyan & Zisserman 2014) and `GoogLeNet` (Szegedy et al. 2015) continued to improve the discriminative power of deep learning algorithms for image classification using deeper and wider NN models, and innovating data augmentation techniques such as scale jittering. Furthermore, `GoogLeNet` provided the means to further improve image classification analysis by introducing multiscale processing, i.e. allowing the NN model to recover local features through smaller convolutions, and abstract features with larger convolutions. In 2015, the `ResNet` (He et al.) model was the first architecture to surpass human performance on the `ImageNet` challenge. In addition to this milestone in computer vision, `ResNet` was also used to demonstrate that a naive stacking of layers does not guarantee enhanced performance in ultradeep NN models, and may actually lead to sub-optimal performance for image classification.

In view of the aforementioned accomplishments, research in deep learning for image classification has become a booming enterprise in science and technology. This vigorous program has led to innovative ways to leverage state-of-the-art NN models to classify disparate data sets. This approach is required because most applications of deep learning for image classification rely on supervised learning. That is, NN models are trained using large data sets of labelled data, such as the `ImageNet` data set. In astronomical research, to enable the morphological classification of galaxies, the deep NN model developed by Dieleman, Willett & Dambre (2015) was trained on $\sim$55 000 galaxy images, each with 40–50 human classifications from the Galaxy Zoo 2 (Willett et al. 2013) online crowdsourcing project. This model was developed for the Galaxy Challenge competition in 2013–14 on the Kaggle platform, and took first place out of 326 entries. Given that data sets of that nature are challenging to obtain, deep 'transfer' learning has provided the means to classify entirely new data sets *by fine-tuning a pre-trained NN model* with the `ImageNet` data set.[1]

While deep transfer learning was initially explored to classify data sets that were of similar nature to those used to train state-of-the-art NN models, the first application of deep transfer learning of a pre-trained `ImageNet` NN model to classify small data sets of entirely different nature was presented in George, Shen & Huerta (2017, 2018), where a variety of NN models were used to report state-of-the-art image classification accuracy of noise anomalies in gravitational wave data. That study triggered a variety of applications of pre-trained `ImageNet` deep learning algorithms to classify images of galactic mergers (Ackermann et al. 2018), and galaxies (Domínguez Sánchez et al. 2018; Barchi et al. 2019; Khan et al. 2019), to mention a few examples.

Building upon these recent successful applications of deep transfer learning for image classification in physics and astronomy, in this paper we demonstrate that deep transfer learning provides the means to classify images of compact star clusters in nearby galaxies obtained with the *Hubble Space Telescope* (*HST*). We show that this approach yields classification accuracies on par with work performed by humans, and has the potential to outperform humans and traditional machine learning. A major motivation of this work is to determine whether these deep transfer learning techniques can be used to automate production-scale classification of candidate star clusters in data from the Cycle 26 *HST*-`Physics at High Angular Resolution in Nearby GalaxieS (PHANGS`[2]) Survey (PI: J.C. Lee, GO-15654) for which observations commenced in 2019 April. *HST*-`PHANGS` is anticipated to yield several tens of thousands of star cluster candidates for classification, only about a half of which will be true clusters. Encoding classification systems in NNs will also improve the consistency of the classifications, and reduce the implicit impacts of subjectivity and subtle differences in classification systems adopted by different individuals (i.e. it can reduce both random and systematic errors in the classifcations).

This paper is organized as follows. In Section 2, we summarize the objectives of star cluster classification, and describe the current classification system, which we employ in this paper. A review of the consistency between human classifications across prior studies is provided to establish the accuracy level to be achieved or surpassed by deep learning in this initial proof-of-concept experiment. In Section 3, we describe the imaging data and classifications used to train our NN models, and then provide an overview of the NN models employed in this work. We report our results in Section 4. We conclude in Section 5 with a summary of the results and next steps for future work.

2 CLASSIFICATION OF COMPACT STAR CLUSTERS IN NEARBY GALAXIES

The objects of interest in this study are compact star clusters and stellar associations in galaxies at distances between 4 and 20 Mpc. The physical sizes of compact clusters are characterized by effective radii between 0.5 pc and about 10 pc (Portegies Zwart, McMillan & Gieles 2010; Ryon et al. 2017). Ryon et al. (2014) report that the distribution of effective radii of young ($\lesssim$10 Myr), massive compact star clusters peaks between 2 and 3 pc based on *HST*

[1] A brief overview of transfer learning is presented in Appendix B.
[2] www.phangs.org

LEGUS observations of NGC 1313 (D ∼ 4 Mpc) and NGC 628 (D ∼ 10 Mpc). Hence, only with the resolution of *HST*[3] can such objects be distinguished from individual stars and separated from other star clusters in galaxies beyond the Local Group. [4] The sizes of stellar associations, which dominate the young stellar population, span a wider range with sizes from a few pc to ∼100 pc (Portegies Zwart et al. 2010; Gouliermis 2018).

Early attempts at classifying clusters in external galaxies with *HST* imaging focused mainly on old globular clusters, for example, the swarm of thousands of globular clusters around the central elliptical galaxy in the Virgo Cluster, M87 (Whitmore et al. 1995). This was a fairly straightforward process since the background was smooth and the clusters were well separated. With the discovery of super star clusters in merging galaxies (e.g. Holtzman et al. 1992), the enterprise of the identification and study of clusters in star-forming galaxies using *HST* began, despite the fact that crowding and variable backgrounds in such galaxies make the process far more challenging. Studies of normal spiral galaxies pushed the limits to fainter and more common clusters (e.g. Larsen 2002; Chandar et al. 2010). In all these early studies, the primary objective was to distinguish true clusters from individual stars and image artefacts, and there were essentially no attempts to further segregate the clusters into different classes.

An exception, and one of the first attempts at a more detailed classification, was performed by Schweizer et al. (1996), who defined nine object types and then grouped them into two classes: candidate globular clusters and extended stellar associations. More recently, Bastian et al. (2012), who studied clusters using *HST* imaging of the M83 galaxy, classified star clusters as either symmetric or asymmetric. Their analysis retained only symmetric clusters, which they posited were more likely to be gravitationally bound. Following this work, many studies in the field, most notably the Legacy ExtraGalactic Ultraviolet (UV) Survey (LEGUS; Calzetti et al. 2015a) began differentiating clusters into two or three different categories, so that they could be studied separately or together depending on the goals of the project (see also the review by Krumholz, McKee & Bland-Hawthorn 2018, and their discussion of 'exclusive' versus 'inclusive' cluster catalogues).

The LEGUS project also employed machine learning techniques for some of their cluster classification work (Messa et al. 2018, Grasha et al. 2019). This pioneering work will be discussed in Section 5.

[3]The WFC3/UVIS point-source function (PSF) full width at half-maximum (FWHM) is 0.″067 at 5000 Å.

[4]We note that for a high-signal-to-noise cluster it is possible to measure the broadening of the image (and hence the size of the source) to a fraction of the FWHM of the PSF of a star. The FWHM of a star using WFC3 is about 1.8 pixel (1.3 pc at D = 4 Mpc, and 6.4 pc at 20 Mpc). A significant amount of testing has been done on ACS and WFC3 images using software like ISHAPE (Larsen 1999), and much published work (including Ryon et al. 2017) has confirmed that this broadening can be measured down to about 0.2 pixel, corresponding to size limits of ∼0.3 pc, ∼0.6 pc at distances of 5 Mpc, 10 Mpc. Extending to 15 and 20 Mpc, the upper end of distance range covered by the PHANGS survey, the cluster size limits are 0.8 and 1.1 pc. Per the ISHAPE manual, at 5 Mpc, this is calculated as 0.2 pixel * 0.04 (arcsec pixel^{-1})* 24 pc arcsec^{-1} * 1.48 = 0.28 pc (where 1.48 is a conversion factor given in the ISHAPE manual when assuming a King profile specifically). Hence, if the peak sizes for clusters are in the 2–3 pc range, the vast majority of cluster will be resolved for most of the galaxies in PHANGS–*HST*.

In LEGUS, cluster candidates are sorted into four classes as follows (Adamo et al. 2017; Cook et al. 2019):

(i) Class 1: compact, symmetric, single central peak, radial profile more extended relative to point source
(ii) Class 2: compact, asymmetric, or non-circular (e.g. elongated), single central peak
(iii) Class 3: asymmetric, multiple peaks, sometimes superimposed on diffuse extended source
(iv) Class 4: not a star cluster (image artefacts, background galaxies, pairs and multiple stars in crowded regions, stars)

We adopt the same classification system for this paper. In general, we refer to class 1, 2, and 3 as 'compact symmetric cluster,' 'compact asymmetric cluster,' and 'compact association', respectively. Examples of objects in each of these classes are shown in Fig. 1.

2.1 Consistency among classifications

The stated goal of this work is to provide cluster classifications via deep transfer learning models that achieve accuracy levels at least as good as other star cluster classifications in the literature, both by human visual inspection and by application of quantitative selection criteria. In this section, we establish this 'accuracy' level, which we define as the consistency between different classifications for the same cluster populations as reported in the literature, as well as relative to classifications homogeneously performed by one of us (Bradley C. Whitmore, hereafter BCW.).

A first look at the overall consistency between the clusters catalogued by different studies, but based on the same data and same limiting magnitude, is provided by the work on M83 by Bastian et al. (2012), Whitmore et al. (2014), and Chandar et al. (2014). Comparisons reported in those papers show that about ∼70 per cent of the clusters are in common between the studies. Later, Adamo et al. (2017) performed a similar comparison for the spiral galaxy NGC 628 for the catalogues from LEGUS and Whitmore et al. (2014), and finds an overlap of ∼75 per cent. Finally, the LEGUS study of M51 by Messa et al. (2018) finds an overlap of 73 per cent in common with a study by Chandar et al. (2016).

These results are not based only upon detailed analysis of human-versus-human cluster classifications for individual objects; they are statistical measures of overlap between samples where a mix of human classification/identification, and automated star/cluster separation based on the concentration index (i.e. the difference in magnitude in a 1 versus 3 pixel radius) were used across the studies.

To more directly evaluate human-versus-human cluster classifications alone, we start with a comparison of the NGC 3351 cluster catalogue from the LEGUS sample (performed by BCW and team member Sean Linden, who was trained by BCW) with a new version of the NGC 3351 cluster catalogue independently constructed by PHANGS–*HST*[5] (performed by BCW alone). This might be viewed as a test of the consistency that might be expected if the same (or very similar) classifiers return to the same data set after a passage of several years. We find an 80 per cent agreement between category 1 objects, 53 per cent for category 2, and 56 per cent for category 3. If

[5]PHANGS–*HST* has expanded imaging coverage of NGC 3351 to produce greater overlap with PHANGS–ALMA CO observations of the galaxy, and is developing new star cluster catalogues for the fields. See Section 3.1 for an overview of the catalogue construction.

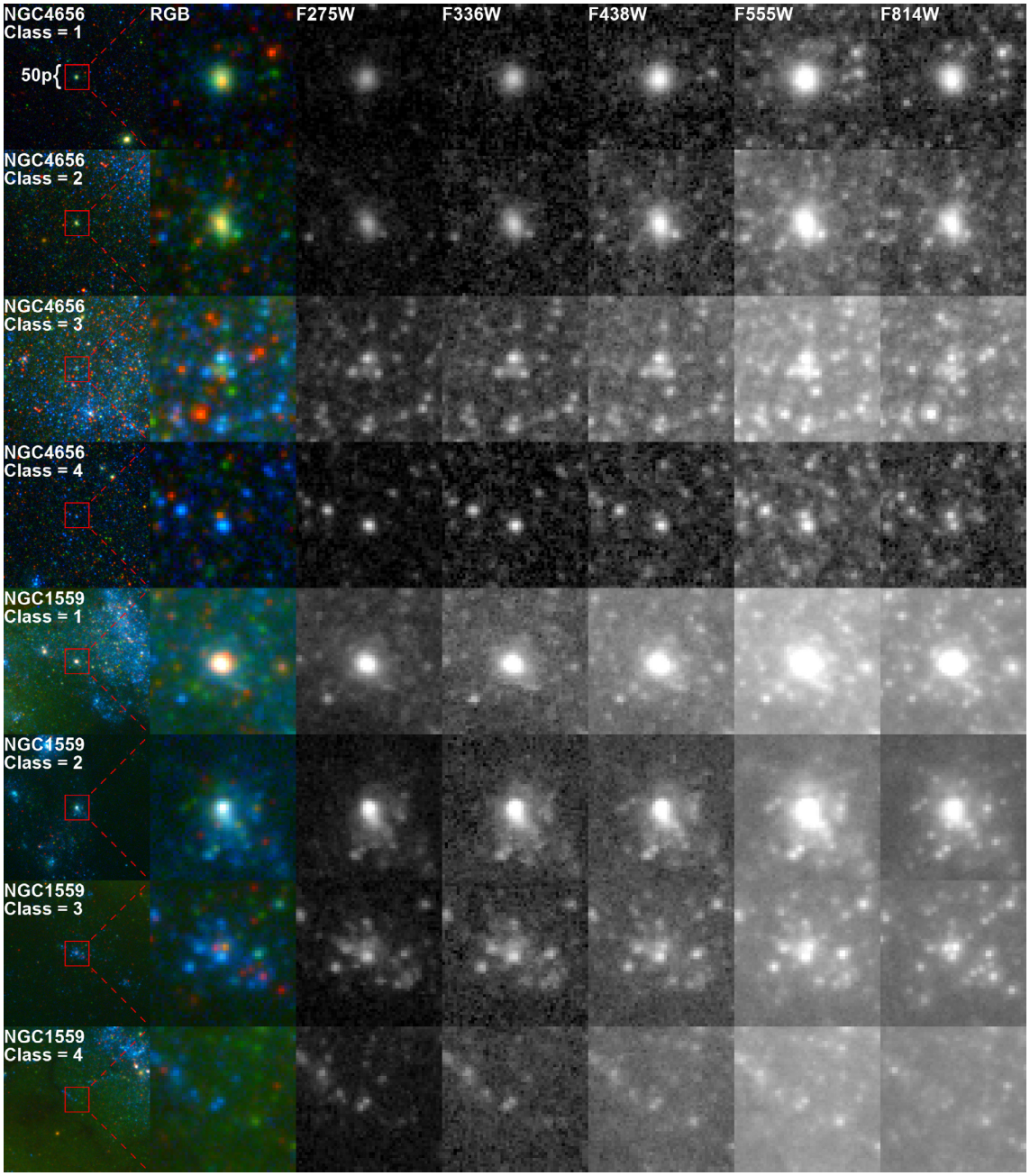

Figure 1. Examples of each of the four cluster classifications illustrated with *HST*/WFC3 imaging. The top four rows show star clusters from NGC 4656, which are part of the training set, while the bottom four rows show clusters from recent PHANGS–*HST* observations of the spiral galaxy NGC 1559, which form our proof-of-concept test sample, and are not used for training. The first two columns show false-colour RGB images for context: the first column displays a 299p × 299p RGB image (R = F814W, G = F438W + F555W, B = F275W + F336W) and the second column shows only the centre 50p × 50p of the RGB image (184pc × 184pc for NGC 1559, for example). The centre 50p × 50p of individual *NUV–U–B–V–I HST* images, which are used as input to the pre-trained NN models for further training (tuning) and evaluation, are shown in the grey scale in the last five columns (from left to right, 50p × 50p images taken with filters F275W, F336W, F438W, F555W, and F814W). We also experiment with 25p × 25p and 100p × 100p images, as discussed in Sections 3 and 4.

we combine category 1 and 2 objects (which is what many authors do for their analysis), the agreement is 88 per cent.

We next compare classifications assigned by BCW for NGC 4656 to those provided in the LEGUS public cluster catalogue, which provides the mode of classifications made by three other LEGUS team members (trained by BCW, A. Adamo, and H. Kim). Results are shown in Fig. 2.

If we combine only the class 1 + 2 clusters (to exclude compact associations that has a higher rate of confusion with class 4 non-clusters), the total match fraction is 67 per cent. For the individual classes, the consistency of the assignments varies from 66 per cent, 37 per cent, 40 per cent, and 61 per cent for class 1, 2, 3, and 4, respectively. Hence, the agreement for the BCW classifications versus the mode of classifications from three LEGUS team members

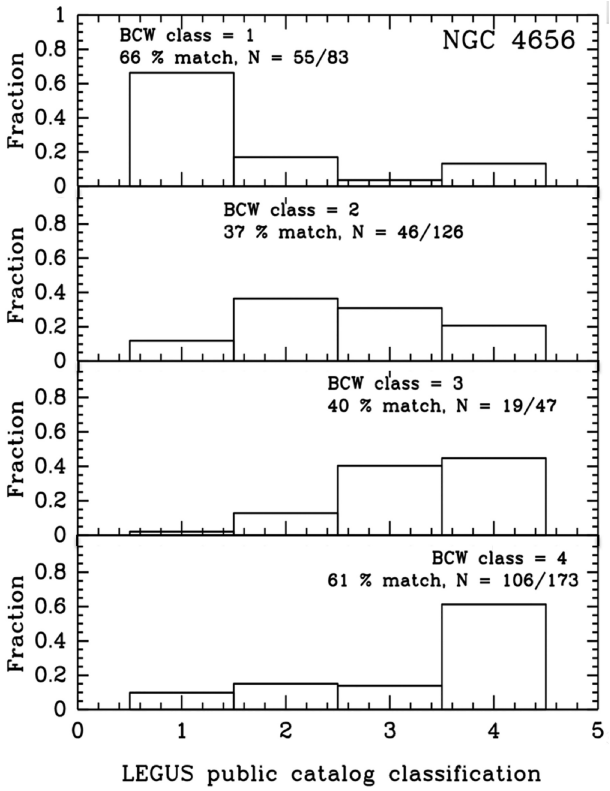

Figure 2. Comparisons between star cluster candidate classifications made by BCW and the mode of classifications made by three other LEGUS team members (trained by BCW, A. Adamo, and H. Kim) provided in the LEGUS public star cluster catalogue for NGC 4656. Each panel shows the distribution of classifications given in the LEGUS catalogue for BCW labelled class 1 (top, symmetric compact clusters), class 2 (upper middle, asymmetric compact clusters), class 3 (lower middle, compact associations), and class 4 (bottom, non-clusters) objects.

for NGC 4656 is slightly lower than the comparisons between BCW and BCW (and Linden) for NGC 3351. Other galaxies where a similar comparison has been made between the BCW classifications and LEGUS 3-person (consensus) classifications (i.e. NGC 4242, NGC 4395N, and M51) result in similar numbers.

In summary, comparing between a wide range of different cluster classification methods, but for the same data sets, we find typical agreements in the range 40 per cent (e.g. when comparing class 2 or class 3 objects alone) to 90 per cent (e.g. when combining class 1 + 2 for repeat classifications of cluster catalogues by the same, or very similar, classifiers). For the individual classes, the 'accuracy' levels that we adopt to be achieved or surpassed for our deep learning studies proof-of-concept demonstration are 70–80 per cent, 40–50 per cent, 40–50 per cent, and 60–70 per cent for class 1, 2, 3, and 4 objects, respectively.

3 DATA AND METHODS

In this section, we describe the data sets used to train, validate, and test our deep learning algorithms, and give an overview of the NN models used. We approach this initial work as a proof of concept demonstration, with the intention of performing further optimization and more detailed tests in future work.

3.1 Star cluster catalogues

A key point is that the training and testing of the NNs presented here are based on a pre-selected sample of cluster candidates where a large fraction of unresolved (point) sources has been first discarded. In past work, such candidate samples have served as the starting point for visual classification by humans to remove remaining interlopers, and to characterize the morphologies of verified clusters as described above. The construction and selection methodolgy for cluster candidate samples used here follow most of the procedures adopted for the LEGUS project (Calzetti et al. 2015b) as described in Adamo et al. (2017).

To briefly review, the procedure includes detection using the SEXTRACTOR program (Bertin & Arnouts 1996) on a white light image; filtering out most stars by requiring the concentration index[6] to be greater than a value determined based on training set of isolated point sources and clusters for each galaxy; requiring detections with photometric errors less than 0.3 mag in at least four filters; and selecting objects brighter than −6 mag in F555W (total Vega magnitude). Again, this results in a cluster candidate list that is then examined visually to remove artefacts (e.g. close pairs of stars, saturated stars and diffraction spikes, background galaxies, etc.). The primary tool used for the visual classification is the IMEXAMINE task in IRAF. See Fig. 3 in Adamo et al. (2017) for a graphic description of the use of IMEXAMINE and the classification into four categories.

For most of the LEGUS star cluster catalogues, which have been publicly released through the MAST archive,[7] classifications are performed by three different team members and the mode is recorded as the final consensus value (i.e. the 29 fields in Table 2). The LEGUS classifiers were trained by BCW, A. Adamo, and H. Kim. For additional eight fields, classifications were performed primarily by a single team member, i.e. coauthor BCW.[8] As of 2019 July, classifications for 4 of the 8 *HST* fields primarily inspected by BCW are available from the LEGUS public archive (Table 1). BCW also independently classified two fields with LEGUS consensus classifications to enable consistency checks (e.g. Fig. 2), bringing the total to 10 galaxies in the sample with BCW classifications.

The construction of a preliminary cluster catalogue for the first galaxy observed in the PHANGS–*HST* program NGC 1559 generally follows the methods used for LEGUS. The primary differences are that a F555W image was used instead of a white light image (which is more prone to small differences in alignment of different filters and the presence of very close pairs of stars with different colours), and a false-colour image from the Hubble Legacy Archive (Whitmore et al. 2016) was simultaneously examined to help classify the clusters. A magnitude limit of −7.5 in the *V* band was used for NGC 1559, reflecting its larger distance (19 Mpc: A. Reiss, private communication) relative to the average distance of the LEGUS galaxies. A detailed presentation of the PHANGS–*HST* star cluster and association candidate selection methods will be provided in the PHANGS–*HST* survey paper (Lee et al., in preparation) and catalogue papers (e.g. Larson et al., in preparation; Thilker et al., in preparation; Whitmore et al., in preparation).

[6](CI = difference in magnitude between an aperture with 1 or 3 pixel).
[7]https://archive.stsci.edu/prepds/legus/dataproducts-public.html
[8]S. Linden, who was trained by BCW, assisted in classifications for sources in NGC 3351, NGC 3627, and NGC 5457).

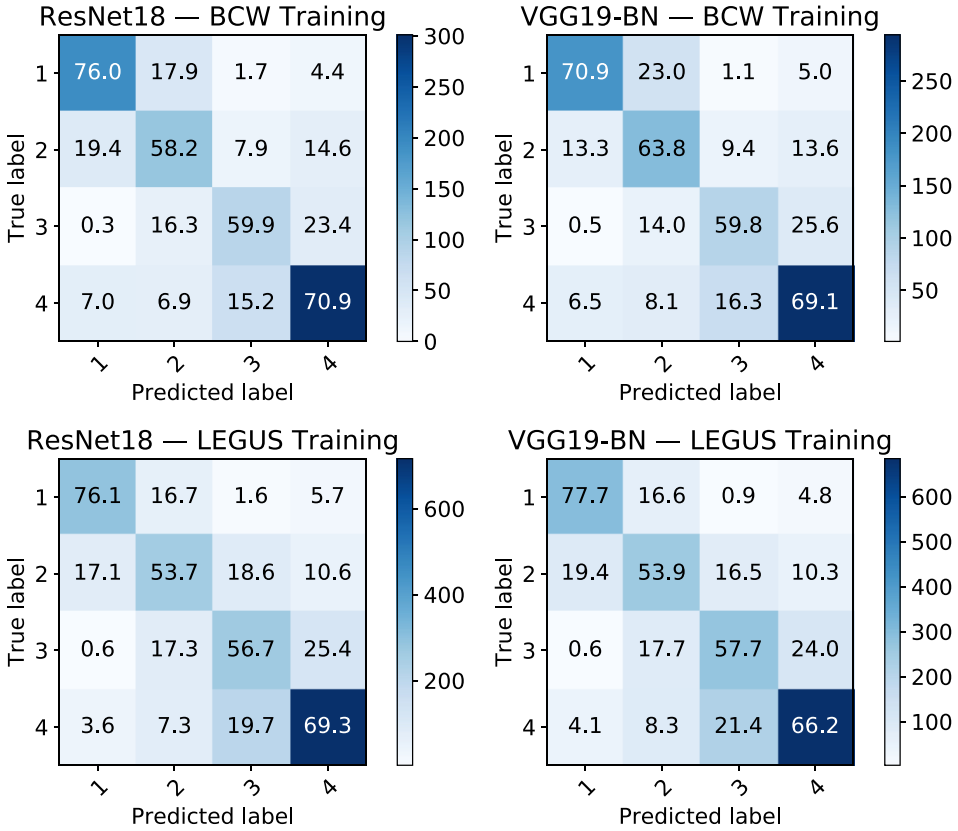

Figure 3. Top panels: Prediction, averaged over 10 models, of `ResNet18` (left) and `VGG19-BN` (right) trained on 80 per cent of the data described in Table 1 and then tested on 20 per cent of the data reserved for validation testing and not used for training. Note that in these confusion matrices each column corresponds to a predicted class, whereas each row corresponds to an actual class. Correct classification results are given along the diagonal from the top left to bottom right of the matrices. The colour bar indicates the number of evaluation images used. Bottom panels: Same as top panels, but for data in Table 2.

3.2 Image data and curation

As input for the NN training, we use postage stamps extracted from *HST* imaging taken in five broad-band filters. Sample postage stamps are presented in the last five columns of Fig. 1.

LEGUS obtained *HST* observations with WFC3 in 2013-2014 (GO-13364; PI Calzetti), and combined those data with ACS data taken in previous cycles by other programs to provide *NUV-U-B-V-I* coverage for a sample of 50 galaxies with 63 fields.

PHANGS–*HST* (GO-15654; PI Lee) began observations on 2019 April 6, and is also obtaining observations with similar exposure times in the *NUV-U-B-V-I* filters. The first galaxy to be observed is NGC 1559.

Bearing in mind that the NN models used in this study (i.e. `VGG19-BN` and `ResNet18`; see next section) were pre-trained with the `ImageNet` data set, in which images are resized to $299 \times 299 \times 3$, we follow best coding practices of NN training, and curate our data sets so that star cluster images have size 299×299 pixels.

However, given that star clusters subtend only about several to a dozen *HST* WFC3 pixels, we focus the training on a small area (see Fig. 1).

We first extract regions of 50×50 *HST*/WFC3 pixels centred on the star cluster candidate, which are then resized to fit in an 299×299 pixel area for the training. With WFC3's pixel size of 0.04 arcsec, each region corresponds to a physical width between $\sim$40 and $\sim$100 pc for our sample of galaxies. To test whether the size of the cropped *HST* image influences the accuracy, we also extract regions that are half and twice as large as 50 *HST*/WFC3 pixels across.

Procedurally, from the *HST* mosaics, a .fits image 'postage stamp' centred on each target cluster is cropped from each of the *NUV–U–B–V–I* bands.

The five resultant stamps for each cluster candidate are then stored in individual header data units within a single multi-extension FITS (multi-extension fits) file. We note that if there was no observation of the cluster in one of the filters, all pixel values for that particular filter's postage stamp were set to zero. If there was no observation in more than one filter, the cluster was removed from our sample, consistent with the candidate selection criteria.

3.3 Neural network models

The available star cluster data sets are small compared to the data sets used to successfully train state-of-the-art NN models for image classification. Thus, we use two NN models, `VGG19` (Simonyan & Zisserman 2014) with batch normalization (`VGG19-BN`) and `ResNet18` (He et al. 2016), pre-trained with the `ImageNet` data set (see Section 1), and then use deep transfer learning[9] to leverage the knowledge of these models to classify real-object

[9]A brief overview of transfer learning is presented in Appendix B.

Table 1. Number of sources in each of the 10 HST LEGUS fields that have been primarily classified by BCW. The number in each of morphological classes described in Section 2 is given. The total number of clusters with detection in at least four filters (a requirement for inclusion in the training and testing) are given in the last row of the table. In total, 80 per cent of the latter (randomly selected) are used for training, and the remaining 20 per cent are reserved for validation testing. Distances compiled by Calzetti et al. (2015a) are listed.

Field	D (Mpc)	Class 1	Class 2	Class 3	Class 4
NGC3351[a]	10.0	118	80	95	325
NGC3627	10.1	403	175	164	837
NGC4242[a]	5.8	117	60	14	42
NGC4395N[b]	4.3	8	19	21	20
NGC4449	4.31	120	261	213	0
NGC45[a]	6.61	45	52	20	43
NGC4656[b]	5.5	83	125	47	173
NGC5457C	6.7	287	108	81	436
NGC5474[a]	6.8	48	95	34	144
NGC6744N	7.1	164	143	58	210
Total		1393	1118	747	2230
N ≥ 4		1271	1013	738	2125

Notes. [a]Classification primarily determined by BCW are available in the public release of the LEGUS cluster catalogues.
[b]Independent classifications determined by BCW for fields for which LEGUS consensus classifications are available through the LEGUS public archive (Table 2).

images to our task at hand, namely, the morphological classification of star clusters.

Regarding batch normalization for VGG19: the weights of each layer in a NN model change throughout the training phase, which implies that the activations of each layer will also change. Given that the activations of any given layer are the inputs to the subsequent layer, this means that the input distribution changes at every step. This is far from ideal because it forces each intermediate layer to continuously adapt to changing inputs. Batch normalization is used to ameliorate this problem by normalizing the activations of each layer. In practice, this is accomplished by adding two trainable parameters to each layer, so the normalized output is multiplied by a standard deviation parameter, and then shifted by a mean parameter. With this approach only two parameters are changed for each activation, as opposed to losing the stability of the network by changing all the weights. It is expected that through this method each layer will learn on a more stable distribution of inputs, which may accelerate the training stage.

Both NN architectures, `VGG19-BN` and `ResNet18` have three input channels. However, since the star cluster candidates have images taken in five broad-band filters, we concatenate two copies of the same NN architecture. The merged NNs have six input channels in total, so we set the input to the last channel to be constant zeros. We also apply one more matrix multiplication and an element-wise softmax function (see Appendix A; Goodfellow, Bengio & Courville 2016) to make sure that for each candidate cluster the output is a vector of size 4, representing the probability distribution over the four classes under consideration. We choose this particular combination given its simplicity and its expected performance for image classification.

We use the pre-trained weights, except those for the last layers, of `VGG19-BN` and `ResNet18` provided by PYTORCH (Paszke et al. 2019) as the initial values for the weights in our models. The weights for the last layers in `VGG19-BN` and `ResNet18` and the last fully

Table 2. Same as Table 1, but for the 29 *HST* LEGUS fields, which have been classified by three people, and have star cluster catalogues available through the LEGUS public archive. The number in each of the morphological classes, as determined by the mode of these three people's classifications, is given.

Field	D (Mpc)	Class 1	Class 2	Class 3	Class 4
NGC1313E	4.39	42	95	122	386
NGC1313W	4.39	85	191	210	373
NGC1433	8.3	51	61	56	138
NGC1566	18.0	258	214	261	328
NGC1705	5.1	16	13	13	54
NGC3344	7.0	119	118	159	161
NGC3738	4.9	49	93	86	214
NGC4656	5.5	93	91	78	169
M51	7.66	363	502	365	1261
NGC5253	3.15	20	37	23	154
NGC628C	9.9	334	357	326	542
NGC628E	9.9	92	80	87	122
NGC6503	5.27	71	96	131	172
NGC7793E	3.44	32	76	83	62
NGC7793W	3.44	51	84	86	78
IC4247	5.1	1	4	3	37
IC559	5.3	9	12	4	18
NGC4395N	4.3	8	12	19	19
NGC4395S	4.3	31	64	42	31
NGC5238	4.51	4	4	1	9
NGC5477	6.4	5	9	9	49
UGC1249	6.9	13	35	40	133
UGC4305	3.05	16	29	40	147
UGC4459	3.66	2	5	3	20
UGC5139	3.98	2	7	7	23
UGC685	4.83	7	4	3	6
UGC695	10.9	4	7	6	94
UGC7408	6.7	19	16	11	32
UGCA281	5.9	2	9	4	34
Total		1799	2325	2278	4866
N ≥ 4		1795	2315	2265	4841

connected layers are randomly initialized. We use cross-entropy as the loss function[10] and `Adam` (Kingma & Ba 2014) for optimization. The learning rate is set to 10^{-4}. The batch size for `ResNet18` is 32, and for `VGG19-BN` is 16.

Batch size and batch normalization refer to two distinct concepts. One epoch corresponds to all the training examples being passed both forward and backward through the NN only once, while the batch size is the number of training examples in one forward/backward pass. For instance, we may divide a training data set of 100 images into four batches, so that the batch size is 25 sample images, and four iterations will complete one epoch. On the other hand, batch normalization is a technique used to improve the stability of the learning algorithms. The details are described in Appendix C.

Finally, following deep learning best practices, we quantify the variance in classification performance of our models by training them 10 times independently and then presenting the mean accuracies and the corresponding standard deviations. We also compute the Shannon entropy Shannon (1948) of the output distribution

[10]A loss function is used to evaluate and diagnose model optimization during training. The penalty for errors in the cross-entropy loss function is logarithmic, i.e. large errors are more strongly penalized.

Table 3. Number of sources in the PHANGS–*HST* observation of NGC 1559 that have been classified by BCW. This cluster sample is used to test the NNs trained as described in Section 4.1 as a proof-of-concept demonstration for production scale classification of PHANGS–*HST* compact clusters and associations.

Field	D (Mpc)	Class 1	Class 2	Class 3	Class 4
NGC1559	19.0	302	252	162	710

over the four star cluster classes to quantify the uncertainty in each individual NN model's prediction.

3.4 Training experiments

We perform a series of experiments to test how the accuracy of the NN model for predicting the morphological classification of candidate star clusters depends on the following characteristics of the training sample:

(i) origin of classifications: primarily classified by BCW (Table 1) or the mode of three LEGUS classifiers (Table 2);
(ii) size of images used for training: 25p × 25p, 50p × 50p, 100p × 100p
(iii) imaging filters: *NUV, U, B, V, I*.

Transfer learning is used to train the NN models using a random selection of 80 per cent of the samples described in Tables 1 and 2 separately, and the remaining 20 per cent is reserved for validation. In total, this results in training samples of about 1000, 800, 600, and 1700 class 1, 2, 3, 4 objects primarily classified by BCW, and about 1400, 1800, 1800, 3900 objects with LEGUS consensus classifications.

Absolute values of pixels are rescaled to be in the range [0, 1], to avoid the brightness of the sources from becoming a parameter in the classification. During training, we use several standard data augmentation strategies, such as random flips, and random rotations in the range $[0,\ 2\pi]$ to make sure that the trained NNs are robust against those transformations. Taking into account the batch size mentioned above for `ResNet18` and `VGG19-BN`, and bearing in mind that we trained the models using about 10 000 batches, this means that the nets were exposed to 320 000 and 160 000 images, respectively. Note, however, that the data augmentation techniques used during the training stage may produce very similar images to the actual star cluster images curated for this analysis.

To investigate whether networks trained in this manner can be used to automate classification of star clusters in the PHANGS–*HST* data set in the future, we test the networks on the first observations obtained by PHANGS–*HST* of the spiral galaxy NGC 1559. The PHANGS–*HST* NGC 1559 observations provide 302, 252, 162, and 710 class 1, 2, 3, 4 objects, as classified by BCW (Table 1).

4 RESULTS

We present four sets of results in this section.

In Section 4.1, we present the classification accuracy for the four categories of star clusters candidates relative to classifications primarily determined by BCW and those based on the mode of classifications performed by three LEGUS team members. We also present the uncertainty quantification analysis of those models (i.e. due to random weight initialization).

In Section 4.2, we quantify the robustness of our NN models to generalize to star cluster images in different galaxies, choosing the PHANGS–*HST* observations of NGC 1559 as the driver of this exercise as discussed above.

In Section 4.3, we report on whether the classification accuracy depends on the size of the images used for network training.

In Section 4.4, we report on relative importance of different filters for image classification in our resulting deep learning models.

4.1 Does prediction accuracy depend on the origin of the classifications?

It is often useful to approach a problem using multiple methods to check how sensitive the results are to the chosen method. For example, the use of both ResNet18 and VGG19-BN architectures in this paper allows us to see which one provides better results, but as we will show next, the results are quite robust no matter which is used. We use a similar strategy in this section by examining the results from training using two different classification samples, namely the BCW sample (see Table 1) and the LEGUS-consensus (three classifiers) sample (see Table 2). While the BCW sample might be expected to have greater internal self-consistency since it was performed by a single experienced classifier, averaging the results of three less-experienced classifiers might be expected to reduce the random noise. Hence, it is not obvious which approach might give better results in this pilot project. In the long run, the development of a much larger standardized data base using a full range of experienced classifiers, as discussed in Section 5, may be required to make significant improvements.

First, we quantify the performance of our models for classification accuracy when we fine-tune the models to determine whether the transfer learning was effective at learning the morphological features that tell apart the four classes of star clusters, and to assess the robustness of the optimization procedure for image classification. As described above, to fine-tune the models pre-trained with the `ImageNet` data set, the weights of the last layers and the last fully connected layers of the `VGG19-BN` and `ResNet18` models are randomly initialized. The process is performed separately for the data sets described in Tables 1 and 2 to examine the dependence of the results on the origin of the classifications.

The results based on training with classifications primarily determined by BCW are presented in the top row of confusion matrices in Fig. 3, for both the ResNet18 and VGG19-BN models, with mean classification accuracy taken as the average over 10 individual trainings from scratch. As a reminder, the reported accuracies are based on classification of a random set of 20 per cent of the overall sample that was not included in the training (the 'validation' sample). Likewise, the results based on training with the mode of classifications performed by three LEGUS team members are presented in the bottom row in Fig. 3.

The main result is that the classification accuracies for the validation samples are comparable for both ResNet18 and VGG19-BN networks, as well as for both training samples. Reading along the diagonal of the confusion matrices presented in Fig. 3, for the models trained on the objects primarily classified by BCW, the accuracies for ResNet18 are 76 per cent, 58 per cent, 60 per cent, and 71 per cent for classes 1, 2 ,3, and 4 objects, respectively, and 71 per cent, 64 per cent, 60 per cent, 69 per cent for VGG19-BN. Similarly, for the networks trained on the mode of classifications performed by three LEGUS members the accuracies are 78 per cent, 54 per cent, 58 per cent, 66 per cent for ResNet18 and 76 per cent, 54 per cent, 57 per cent, 69 per cent for VGG19-BN. This provides evidence that our proof-of-concept NN models are resilient to the

Table 4. Prediction of `ResNet18` on 20 per cent of the data in Table 1 reserved for validation testing and not included in the training, averaged over 10 models. The averaged predictions from Fig. 3 are repeated, but now the standard deviations are also shown. The number of validation images for each class are listed in the final column.

	Class 1 (per cent)	Class 2 (per cent)	Class 3 (per cent)	Class 4 (per cent)	**Total**
BCW Class 1	**76.0 ± 4.2**	17.9 ± 4.4	1.7 ± 0.7	4.4 ± 1.4	254
BCW Class 2	19.4 ± 3.5	**58.2 ± 5.3**	7.9 ± 3.5	14.6 ± 3.0	202
BCW Class 3	0.3 ± 0.5	16.3 ± 5.4	**59.9 ± 6.8**	23.4 ± 5.6	147
BCW Class 4	7.0 ± 2.1	6.9 ± 2.9	15.2 ± 3.1	**70.9 ± 4.8**	425

Table 5. As Table 4, but now using `VGG19-BN`.

	Class 1 (per cent)	Class 2 (per cent)	Class 3 (per cent)	Class 4 (per cent)	Total
BCW Class 1	**70.9 ± 6.2**	23.0 ± 4.8	1.1 ± 0.7	5.0 ± 1.9	254
BCW Class 2	13.3 ± 4.3	**63.8 ± 4.8**	9.4 ± 2.9	13.6 ± 3.6	202
BCW Class 3	0.5 ± 0.7	14.0 ± 6.3	**59.8 ± 7.5**	25.6 ± 7.4	147
BCW Class 4	6.5 ± 2.4	8.1 ± 2.6	16.3 ± 3.8	**69.1 ± 6.8**	425

choice of data used for training and validation despite the fact that the two samples were (i) labelled by different classifiers; and (ii) include different parent galaxies at a wide range of distances (4–10 Mpc for the objects primarily classified by BCW, and 4–18 Mpc for the sample with LEGUS consensus classifications.) Our findings indicate that notwithstanding these seemingly important differences, the prediction accuracies using these two independent data sets are fairly consistent.

The variance in the 10 independent classification measurements provide measure of the robustness of the models. The variances for our NN models trained on the classifications primarily determined by BCW are given in Tables 4 and 5. In all cases, the variances are between 4 and 8 per cent. The variances for LEGUS classifications are comparable.

4.2 How accurately can the models predict classifications for clusters in galaxies not included in the training sample?

To further assess the robustness and resilience of our NN models, we use them to classify images from a galaxy not included in the original training data set, namely the PHANGS–*HST* target NGC 1559 (see Table 3). This galaxy is about two to four times further away than the galaxies in either of the training samples, with the notable exception of NGC 1566, which is at a comparable distance to NGC 1559 (18 versus 19 Mpc), and included the sample with consensus classifications from three LEGUS team members (Table 2). Results are presented in Fig. 4 and Tables 6 and 7.

Notwithstanding these differences, we again notice that all models produce comparable results. Reading along the diagonal of the confusion matrices presented in Fig. 4, for the models trained on the objects primarily classified by BCW, the accuracies for ResNet18 are 73 per cent, 38 per cent, 40 per cent, 75 per cent for class 1, 2 ,3, and 4 objects, respectively, and 74 per cent, 42 per cent, 52 per cent, 67 per cent for VGG19-BN. Likewise, for the networks trained on the mode of classifications performed by three LEGUS members the accuracies are 70 per cent, 41 per cent, 48 per cent, 62 per cent for ResNet18 and 70 per cent, 45 per cent, 52 per cent, 52 per cent for VGG19-BN.

For all models, the performance for NGC 1559 class 1 star clusters is at or above the 70 per cent level. The classification accuracy of the BCW-based models is similar to their performance on the validation samples (i.e. Fig. 3). Meanwhile for NGC 1559 class 1 star clusters the performance of the models trained on the LEGUS consensus classifications are 6–8 per cent lower relative to the classification of the validation samples. On the other hand for class 2 star clusters, the accuracies hover around the 40 per cent level, and are the lowest of the four classes. The accuracies for the models trained on the objects primarily classified by BCW drop by ∼20 per cent: from 58 per cent (test subset sample) to 38 per cent (NGC 1559) for ResNet18, and from 64 per cent to 42 per cent for VGG19-BN. Similarly, those trained on the LEGUS consensus classifications drop, although by only ∼10 per cent: from 54 per cent to 41 per cent for ResNet18, and from 54 per cent to 45 per cent for VGG19-BN. The accuracies for the NGC 1559 class three star clusters are at the 40–50 per cent level, a ∼10 per cent drop for all models relative to the performance on the test subsets. Finally, for the class 4 non-clusters, the models trained on the objects primarily classified by BCW perform comparably, i.e. at the 70 per cent level, while those trained on the LEGUS consensus classifications drop to the 50–60 per cent level.

4.2.1 Uncertainty calculations through entropy analysis

Another method to investigate the uncertainty in the models' predictions is through the computation of entropy using the probability distributions for each of the cluster classes we are trying to classify, which is an output of the models. Intuitively, the more pronounced the peak is in the probability distribution, the more confident the NN is about its prediction, and in this case, the entropy calculated from the prediction probability distribution will be lower. For example, if the probability distribution is only concentrated on one class, the network network in this case is 100 per cent certain about its prediction and the entropy would be zero, i.e. there is no uncertainty. On the other hand, if the prediction assigned the same probability for all the for classes under consideration equally, we would have maximum uncertainty in this case, since for the given input image, all the four classes are equally possible to be the predicted classes,

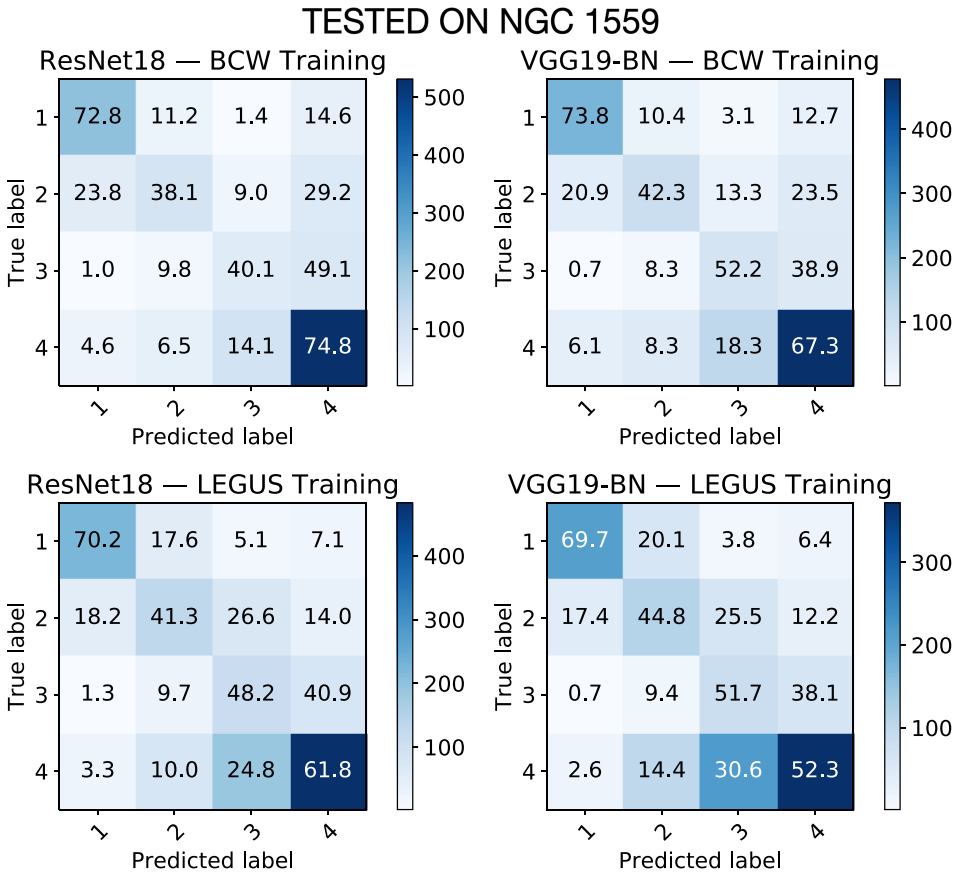

Figure 4. Top panels: Same as Fig. 3, but now the models trained on the classifications primarily determined by BCW (Table 1) are applied to predict classifications for candidates in PHANGS–*HST* observations of NGC 1559, a galaxy that was not included in the training samples. As before, results were obtained after averaging over 10 models. Bottom panels: Same as top row, but for models trained on the mode of classifications performed by three LEGUS team members (Table 2).

Table 6. Prediction of `ResNet18` trained on star clusters primarily classified by BCW (Table 1) for candidates in spiral galaxy NGC 1559 from the PHANGS–*HST* program, averaged over 10 models. Each row shows the averaged predictions (same as shown in top left-hand panel of Fig. 4), but now together with the standard deviations from the 10 models. The number of objects classified is given in the last column. This experiment was performed to test the ability of this neural network model to generalize to images from galaxies not included in the training sample. It is notable that NGC 1559 is roughly twice as far away as any of galaxies in the BCW training sample.

	Class 1 (per cent)	Class 2 (per cent)	Class 3 (per cent)	Class 4 (per cent)	Total
BCW Class 1	**72.8 ± 7.6**	11.2 ± 3.8	1.4 ± 0.6	14.6 ± 5.1	302
BCW Class 2	23.8 ± 4.3	**38.1 ± 5.9**	9.0 ± 4.0	29.2 ± 4.7	252
BCW Class 3	1.0 ± 0.5	9.8 ± 4.2	**40.1 ± 7.1**	49.1 ± 6.1	162
BCW Class 4	4.6 ± 1.4	6.5 ± 1.8	14.1 ± 3.1	**74.8 ± 3.5**	710

Table 7. As Table 6, but now using `VGG19-BN`.

	Class 1 (per cent)	Class 2 (per cent)	Class 3 (per cent)	Class 4 (per cent)	Total
BCW Class 1	**73.8 ± 4.8**	10.4 ± 3.5	3.1 ± 1.3	12.7 ± 4.4	302
BCW Class 2	20.9 ± 6.4	**42.3 ± 7.9**	13.3 ± 2.6	23.5 ± 8.0	252
BCW Class 3	0.7 ± 0.6	8.3 ± 3.3	**52.2 ± 5.9**	38.9 ± 7.5	162
BCW Class 4	6.1 ± 2.4	8.3 ± 3.3	18.3 ± 3.0	**67.3 ± 6.8**	710

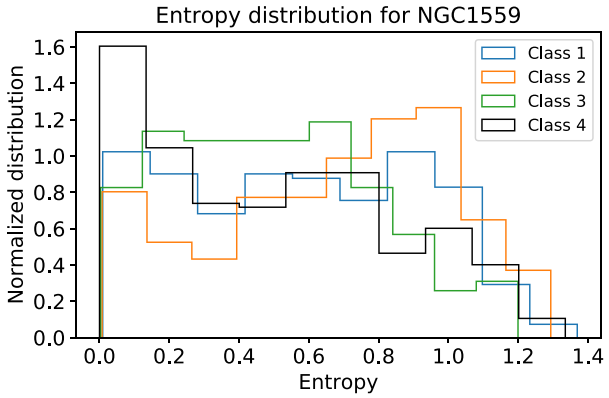

Figure 5. The uncertainty in the neural network's prediction is quantified by the entropy of the predicted probability distribution over the four star cluster image classes considered in this analysis. For a random guess over the four classes, the entropy is $\ln(4) \approx 1.39$. The lower the entropy, the higher the confidence the neural network has about its prediction. The panel shows the predicted entropy value for each NGC 1559 image with which we classified with our `VGG19-BN` model, trained on the objects primarily classified by BCW given in Table 1. The *x*-axis shows the binned values of the entropy values, whose frequency of occurrence is indicated on the *y*-axis. To make clear that the area of each histogram is normalized to one, the *y*-axis label is explicitly labelled 'Normalized distribution.'.

and in this case, the maximum entropy is $\ln(4) \approx 1.39$. Fig. 5 shows the distribution of the entropies for the predictions of `VGG19-BN` when tested on NGC 1559 images.

4.3 How does classification accuracy depend on size of training images?

To quantify the importance of image size for star cluster classification, we train our NN models again, but with two additional cropping sizes: 25 × 25 pixels and 100 × 100 pixels. In Fig. 6, we present results from training on the sample with LEGUS consensus classifications (again, where 80 per cent of the sample is used for training and 20 per cent for testing), where the results presented earlier from our fiducial experiments with 50 × 50 pixels postage stamps are repeated to facilitate comparison. We present results based on the LEGUS consensus classifications as the range of distances of the galaxies (from 3.1 to 18 Mpc; Table 2) is inclusive of the range spanned by the sample primarily classified by BCW (Table 1). Hence, the physical scales subtended by the cropped images span from 16 pc (for 25 × 25 pixel images at 3.1 Mpc) to 360 pc (for 100 × 100 pixel images at 18 Mpc).

There are no significant differences between the results for the different cropping sizes. These results indicate that our NN models are resilient to this particular data curation choice. We see variations at the level of ∼ 5 per cent, which is within the expected variation in the performance of the NN models due to random weight initialization, as indicated in Tables 4 and 5. Results for the models trained with objects primarily classified by BCW are consistent. The results also do not change if the NN models are trained with postage stamps using *random* cropping sizes ranging from 25 × 25 pixels to 100 × 100 pixels (i.e. a random cropping size is chosen for each object in the training and testing sample).

4.4 Classification accuracy as a function of imaging filter

We have also quantified what filter has the leading contribution for classification accuracy. To do so, we perform the following experiment: using NGC 1559 images as testing data set, we produced five different testing data sets in which one filter was set to zero. We then fed these five different testing data sets, one at a time, to our NN models trained with objects primarily classified by BCW and quantified which missing filter leads to the most significant drop in classification accuracy. As shown in Fig. 7, the key filter is F555W.

This finding is expected, since the human classifications primarily rely on the F555W image (e.g. using DS9 and imexamine), with colour images (F814, F555, F336W) generated by the Hubble Legacy Archive providing supporting morphological information. Therefore, our NN models seem to use insights similar to human vision to classify star cluster images.

5 DISCUSSION AND CONCLUSIONS

Using homogeneous data sets of human-labelled star cluster images from the *HST*, we have leveraged a new generation of NN models and deep transfer learning techniques for morphological classification of compact star clusters in nearby galaxies to distances of ∼ 20 Mpc. These results are very promising.

(i) Through all of the experiments presented here with multiple training sweeps for each NN model, we see that the classification accuracy is similar for both architectures studied, i.e. ResNet18 and VGG19-BN pre-trained with the ImageNet data set where the weights of the last layers and the last fully connected layers are randomly initialized.

(ii) Somewhat surprisingly, the performance of the models is relatively robust to the origin of the human classifications used, the particular galaxies included in the training sample, and the cropping size of the training images (spanning physical sizes of 16–360 pc). Irrespective of whether the models are trained on a sample primarily classified by one expert (BCW) with galaxies at distances 2–4 times closer than the star cluster candidates to be evaluated in PHANGS–*HST* galaxy NGC 1559; or trained on the mode of classifications from three individuals where the sample does include a galaxy at a distance similar to NGC 1559; the results are comparable. The prediction accuracies for NGC 1559, which was not included in the training samples, are at the level of 70 per cent:40 per cent 40–50 per cent for the class 1, 2, and 3 star clusters. However, the BCW-trained networks have a higher performance in classification of the class 4 non-clusters in NGC 1559 (70 per cent versus 50–60 per cent). This might be expected since the classifications for NGC 1559 were also performed by BCW, and may be due to a higher level of self-consistency in the training and testing classification data sets.

(iii) Most importantly, despite training with relatively small data sets, the performance of the networks presented here is competitive with the consistency achieved in previous human and quantitative automated classification of the same star cluster candidate samples (Section 2.1). Thus, this work provides a proof-of-concept demonstration that deep transfer learning can be successfully used to automate morphological classification of star cluster candidate samples using *HST* UV-optical imaging being obtained by PHANGS–*HST*.

This work represents a milestone in the use of deep transfer learning for this area of research, and represents progress from initial machine learning experiments described in Grasha et al.

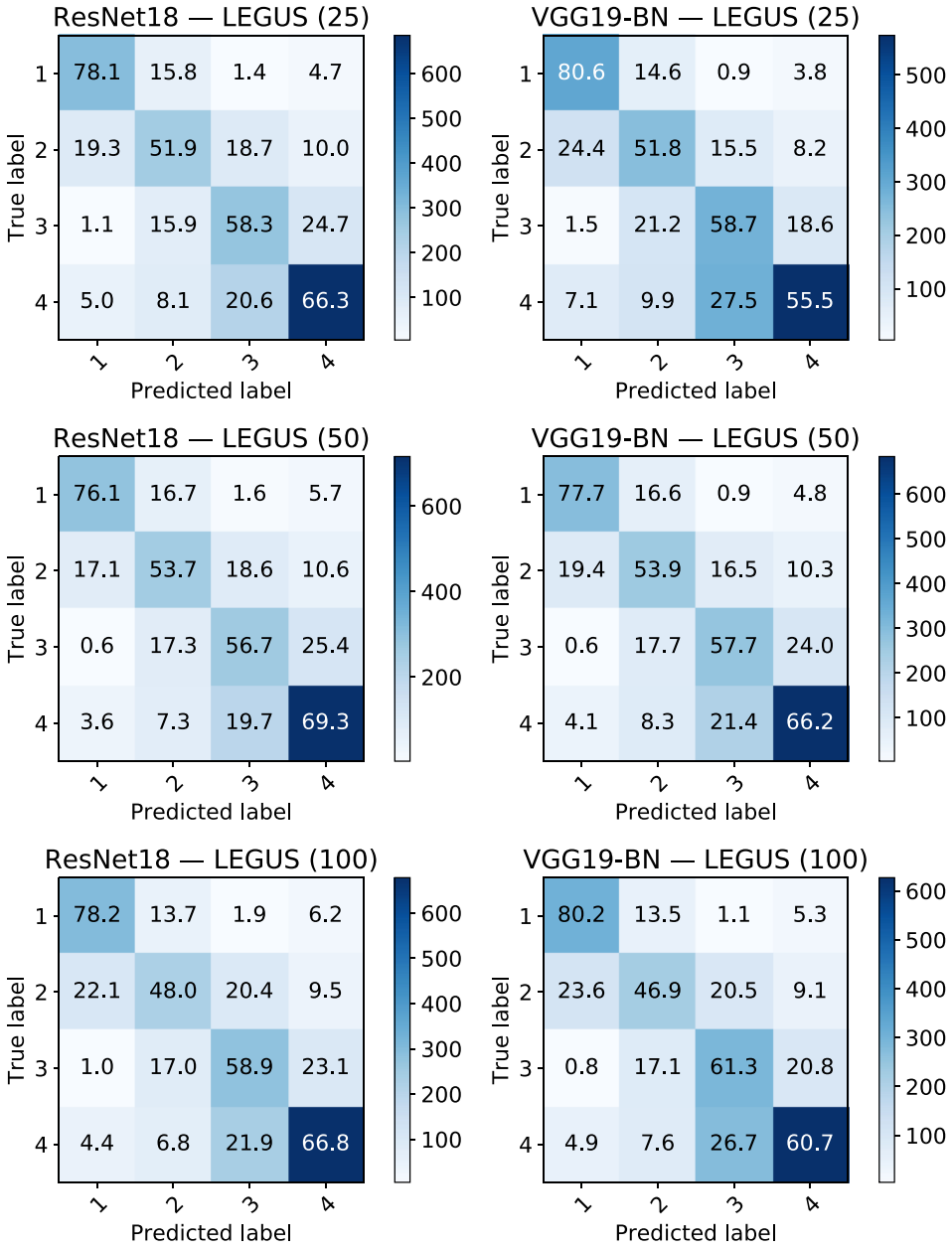

Figure 6. Left column: `VGG19-BN` model classification results for cropping size 25 × 25, 50 × 50 and 100 × 100. Right column: as before, but now for `ResNet`.

(2019) and also discussed in Messa et al. (2018). Grasha et al. (2019) experimented with the use of an ML algorithm for classifying the approximately 11 000 clusters in the spiral galaxy M51, based on a human classified training set with ∼2500 clusters from the LEGUS sample. While the recovery of class 1 and 2 clusters is fairly good (in the range 60–75 per cent in the Grasha and Messa studies, and comparable to the prediction accuracies presented here) recovery of class 3 clusters is poor, with an apparently significant anticorrelation.

To attempt to further improve upon the models presented here, future work will include training with the largest star cluster candidate sample possible (i.e. combining all samples used for this proof-of-concept demonstration plus classifications for objects in several galaxies in PHANGS–*HST*). Improvement in classification accuracy also requires the development of a standarized data set of human-labelled star cluster classifications, with classifications agreed upon by a full range of experts in the field, to be used as the basis for future network training. This effort would benefit from a classification challenge, where experts can come to detailed agreement on the morphological features that constitute the criteria for classification (e.g. to establish full decision trees, such as those used for Galaxy Zoo by citizen scientists), and explicitly describe where they disagree and why. A review of differences in star cluster definitions between research groups, and their possible impact on conclusions about star cluster formation and evolution, can be found in Krumholz et al. (2018). The ultimate goal is to use deep learning

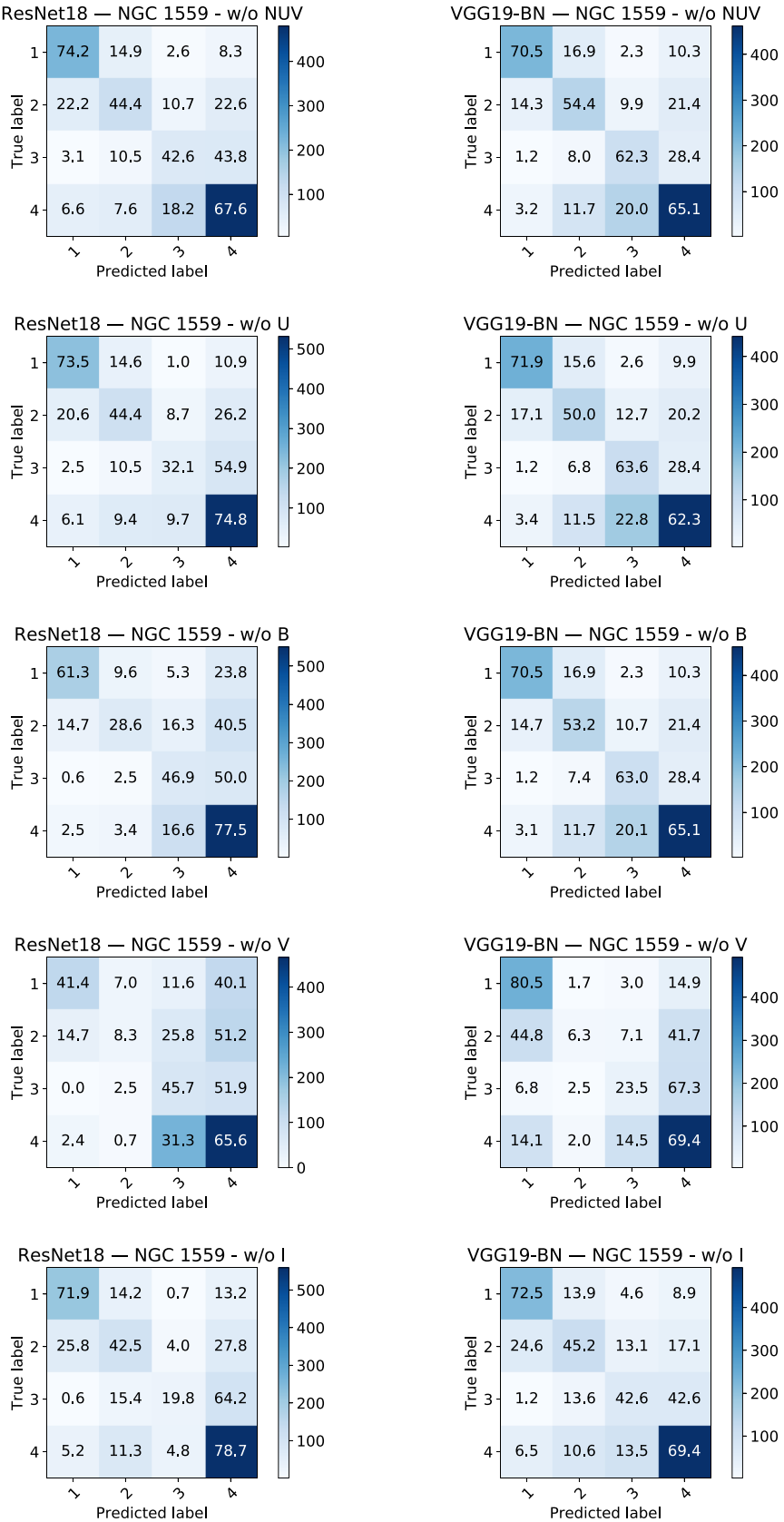

Figure 7. Left column: `ResNet` model classification results when the indicated filter is removed from the composite image. Right column: as before, but now for `VGG19-BN`. The greatest drop in the accuracies occurs when the *V*-band filter is removed.

techniques to not only rapidly produce reliable classifications and speed the time to science, but to significantly advance the field of star cluster evolution. Given the discussion in Krumholz et al. (2018), this requires that deep learning networks are trained on such standardized data sets, broadly adopted by workers in the field.

With this study, we open a new chapter to explore in earnest the use of deep transfer learning for the classification of very large data sets of star cluster galaxies in ongoing and future electromagnetic surveys, and application to the new PHANGS–*HST* data being obtained now.

ACKNOWLEDGEMENTS

We thank the referee for feedback that significantly improved the paper, and in particular, motivated the expansion of our experiments to investigate potential differences in outcome when training with classifications by a single individual (BCW) versus using LEGUS consensus classifications from multiple individuals. Initially, our experiments were based solely on the classifications of BCW.

We also thank the LEGUS team, and in particular Daniela Calzetti and Kathryn Grasha, for their pioneering efforts in the field of classifying star clusters using machine learning techniques, and for making results available via the LEGUS public website. We thank Sean Linden for assisting BCW in classifications for star cluster candidates in NGC 3351, NGC 3627, and NGC 5457.

Based on observations made with the NASA/ESA *HST*, obtained from the data archive at the Space Telescope Science Institute. STScI is operated by the Association of Universities for Research in Astronomy, Inc. under the National Aeronautics and Space Administration (NASA) contract NAS 5-26555. Support for Program number 15654 was provided through a grant from the STScI under NASA contract NAS5- 26555.

This research has used the NASA/IPAC Extragalactic Database, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

EAH and WW gratefully acknowledge National Science Foundation (NSF) awards OAC-1931561 and OAC-1934757.

This research is part of the Blue Waters sustained-petascale computing project, which is supported by NSF awards OCI-0725070 and ACI-1238993, and the State of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

This work utilized resources supported by the NSF's Major Research Instrumentation program, grant OAC-1725729, as well as the University of Illinois at Urbana-Champaign.

We are grateful to NVIDIA for donating several Tesla P100 and V100 GPUs that we used for our analysis, and the NSF grants NSF-1550514, NSF-1659702, and TG-PHY160053.

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. We thank the NCSA Gravity Group for useful feedback.

MC and JMDK gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) through an Emmy Noether Research Group (grant KR4801/1-1) and the DFG Sachbeihilfe (grant KR4801/2-1). JMDK gratefully acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the ERC Starting grant MUSTANG (grant 714907).

REFERENCES

Abbott T. et al., 2016, MNRAS, 460, 1270
Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415
Adamo A. et al., 2017, ApJ, 841, 131
Ball N. M., Brunner R. J., Myers A. D., Tcheng D., 2006, ApJ, 650, 497
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12
Banerji M. et al., 2010, MNRAS, 406, 342
Barchi P. H., de Carvalho R. R., Rosa R. R., Sautter R., Soares-Santos M., Marques B. A. D., Clua E., 2019, Astron. Comput., 30, 100334
Bastian N. et al., 2012, MNRAS, 419, 2606
Bengio Y., 2011, AIP Conf. Proc. Vol. 27, Quantum Communication, Measurement and Computing (QCMC): The Tenth International Conference. Am. Inst. Phys., New York, p. 17
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Calzetti D. et al., 2015a, AJ, 149, 51
Calzetti D. et al., 2015b, ApJ, 811, 75
Cannon A. J., Pickering E. C., 1912, Ann. Harv. Coll. Obs., 56, 65
Cannon A. J., Pickering E. C., 1918, Ann. Harv. Coll. Obs., 91, 1
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Chandar R. et al., 2010, ApJ, 719, 966
Chandar R., Whitmore B. C., Calzetti D., O'Connell R., 2014, ApJ, 787, 17
Chandar R., Whitmore B. C., Dinino D., Kennicutt R. C., Chien L.-H., Schinnerer E., Meidt S., 2016, ApJ, 824, 71
Cook D. O. et al., 2019, MNRAS, 484, 4897
de Vaucouleurs G., 1963, ApJS, 8, 31
Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in CVPR09, Available at: http://www.image-net.org/
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Domínguez Sánchez H. et al., 2018, MNRAS, 484, 93
Everingham M., Eslami S. M. A., Van Gool L., Williams C. K. I., Winn J., Zisserman A., 2015, Int. J. Comput. Vis., 111, 98
Fadely R., Hogg D. W., Willman B., 2012, ApJ, 760, 15
George D., Shen H., Huerta E. A., 2017, Phys. Rev. D, 97, 101501(R)
George D., Shen H., Huerta E. A., 2018, Phys. Rev. D, 97, 101501
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA
Gouliermis D. A., 2018, PASP, 130, 072001
Grasha K. et al., 2019, MNRAS, 483, 4707
He K., Zhang X., Ren S., Sun J., 2016, Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770, Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
Holtzman J. A. et al., 1992, AJ, 103, 691
Hubble E. P., 1926, ApJ, 64, 321
Hubble E. P., 1936, Realm of the Nebulae. Yale Univ. Press, New Haven
Ishak B., 2017, Contemp. Phys., 58, 99
Kamdar H. M., Turk M. J., Brunner R. J., 2016, MNRAS, 455, 642
Khan A., Huerta E. A., Wang S., Gruendl R., Jennings E., Zheng H., 2019, Phys. Lett., B795, 248
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
Krizhevsky A., Sutskever I., Hinton G. E., 2012, Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 1. NIPS'12. Curran Associates Inc., USA, p. 1097. Available at: http://dl.acm.org/citation.cfm?id = 2999134.2999257
Krumholz M. R., McKee C. F., Bland-Hawthorn J., 2018, ARA&A, 57, 227
Larsen S. S., 1999, Astronomy and Astrophysics Supplement, 139, 393
Larsen S. S., 2002, AJ, 124, 1393
Lecun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Małek K. et al., 2013, A&A, 557, A16
Messa M. et al., 2018, MNRAS, 473, 996
Paszke A. et al., 2019, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., p. 8024

Portegies Zwart S. F., McMillan S. L. W., Gieles M., 2010, ARA&A, 48, 431
Russakovsky O. et al., 2015, Int. J. Comput. Vis., 115, 211
Ryon J. E. et al., 2014, AJ, 148, 33
Ryon J. E. et al., 2017, ApJ, 841, 92
Schweizer F., Miller B. W., Whitmore B. C., Fall S. M., 1996, AJ, 112, 1839
Sevilla-Noarbe I., Etayo-Sotos P., 2015, Astron. Comput., 11, 64
Shannon C. E., 1948, Bell Syst. Tech. J., 27, 379
Simonyan K., Zisserman A., 2014, CoRR, abs/1409.1556, preprint (arXiv: 1409.1556)
Solarz A., Bilicki M., Gromadzki M., Pollo A., Durkalec A., Wypych M., 2017, A&A, 606, A39
Suchkov A. A., Hanisch R. J., Margon B., 2005, AJ, 130, 2439
Szegedy C. et al., 2015, Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 1, preprint (arXiv:1409.4842)
Vasconcellos E. C., de Carvalho R. R., Gal R. R., LaBarbera F. L., Capelato H. V., Frago Campos Velho H., Trevisan M., Ruiz R. S. R., 2011, AJ, 141, 189
Weir N., Fayyad U. M., Djorgovski S., 1995, AJ, 109, 2401
Whitmore B. C., Sparks W. B., Lucas R. A., Macchetto F. D., Biretta J. A., 1995, ApJ, 454, L73
Whitmore B. C. et al., 2014, ApJ, 795, 156
Whitmore B. C. et al., 2016, AJ, 151, 134
Willett K. W. et al., 2013, MNRAS, 435, 2835

APPENDIX A: STATISTICAL FOUNDATIONS OF DEEP LEARNING CLASSIFIERS

Within the framework of statistical learning, an image X can be modelled as a random matrix that takes value in set $\mathcal{X}$, and the corresponding class can be treated as a random variable Y that takes value in set $\mathcal{Y}$. Since we use 299×299 images with five channels, we treat a cluster image as random matrix of size $299 \times 299 \times 5$. Similarly, as we are trying to classify the images into four classes, Y is a discrete random variable that takes values in $\mathcal{Y}$ with cardinality $|\mathcal{Y}| = 4$.

We assume that the star images and the corresponding class labels follow some unknown but fixed joint probability distribution, with the probability density function $f_{XY}(x, y)$. We also use $\Delta_{\mathcal{Y}}$ to denote set of all possible distribution over $\mathcal{Y}$. Since in our case, $|\mathcal{Y}| = 4$, we have $\Delta_{\mathcal{Y}} = \{\pi = (\pi_1, \pi_2, \pi_3, \pi_4) : \sum_{i=1}^{4} \pi_i = 1, \pi_i \geq 0, \forall i \in [4]\}$.

Under these conventions, the goal of classification is to find a classifier or function $h : X \rightarrow \Delta_{\mathcal{Y}}$ that minimizes the expectation of the cross entropy between the predicted and the ground truth probability mass distribution (pmf) over the classes given the input image X, namely,

$$L(h) = \mathbf{E}[H(h(X),\, f_{Y|X}(\cdot|X))] \tag{A1}$$

$$= \int H(h(X),\, f_{Y|X}(\cdot|x)) f_X(x) \mathrm{d}x, \tag{A2}$$

where $f_X(x)$ is the marginal distribution of X over $\mathcal{X}$, and H is the cross entropy between the predicted and the ground truth pmf over classes,

$$H(h(x),\, f_{Y|Y}(\cdot|x)) = -\sum_{i=1}^{4} f_{Y|X}(Y = i|x) \log([h(x)]_i), \tag{A3}$$

and the $f_{Y|X}(y|x)$ is the conditional distribution of Y given X.

In most cases, we only know the empirical distribution $\hat{f}_{XY}(x, y)$ of (X, Y) and $\hat{f}_{Y|X}(y|x)$ of Y, which are determined by the empirical data. So, the quantity we can directly minimize is

$$\hat{L}(h) = \hat{\mathbf{E}}[H(h_X(\cdot),\, \hat{f}_{Y|X}(\cdot|X))] \tag{A4}$$

$$= \int H(h_x(\cdot),\, \hat{f}_{Y|X}(\cdot|x)) \hat{f}_X(x) \mathrm{d}x\,, \tag{A5}$$

In practice, if the choice of $h(\cdot)$ is arbitrary, then finding an optimal solution is computationally unfeasible. Therefore, we often restrict the searching space to a class of parametrized functions, $h_{\boldsymbol{w}}(\cdot)$, where $\boldsymbol{w}$ is a vector of parameters. In this case, the optimization problem can be posed as

$$\boldsymbol{w}^* = \mathrm{argmin}_{\boldsymbol{w}} \hat{L}(h_{\boldsymbol{w}})\,. \tag{A6}$$

The choice of the parametrized function class is critical to the success of any statistical learning algorithm. In recent years, a deep-layered structure of functions has received much attention (LeCun, Bengio & Hinton 2015; Goodfellow et al. 2016):

$$h_{\boldsymbol{w}}(\mathbf{x}) = h_{\boldsymbol{w}_n}(h_{\boldsymbol{w}_{n-1}}(\cdots h_{\boldsymbol{w}_1}(\mathbf{x}))), \tag{A7}$$

where n is the number of layers or the depth. Usually, we choose, $h_{\boldsymbol{w}_i}(\boldsymbol{x}) = g(\boldsymbol{w}_i \boldsymbol{x})$, where $\boldsymbol{w}_i$ is a matrix, $\boldsymbol{x}$ is an input vector, and $g(\cdot)$ is a fixed non-linear function, e.g. $\max\{\cdot, 0\}$ (also known as `ReLU`), $\tanh(\cdot)$, etc., which is applied element-wise. For the classification problems, we usually apply the so-called softmax function after the last linear transformation. The softmax function on a vector $\boldsymbol{x}$ is a normalization after an element-wise exponentiation:

$$\mathrm{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{i=1}^{n} \exp(x_i)}, \quad \forall i = 1, ..., n, \tag{A8}$$

where n is the length of $\boldsymbol{x}$.

This function class and its extensions, also dubbed NNs, combined with simple first-order optimization algorithms such as stochastic gradient descent (SGD), and improved computing hardware, has lead to disruptive applications of deep learning (LeCun et al. 2015; Goodfellow et al. 2016).

APPENDIX B: DEEP TRANSFER LEARNING

In practice, equation (A6) is usually iteratively solved using variants of SGD. Thus, the choice of initial value for weights $\boldsymbol{w}$ is critical to the success of the training algorithm. If we have some prior knowledge about what initial wights m $\boldsymbol{w}_0$ works better, then it is highly possible that the numerical iteration can converge faster and return better weights $\boldsymbol{w}$. This is the idea behind deep transfer learning (Bengio 2011; Goodfellow et al. 2016).

For a deep learning NN, such as the one defined by equation (A7), the layered structure can be intuitively interpreted as different levels of abstraction for the learned features. In other words, layers that are close to the input learn lower level features, such as different shapes and curves in the image, and layers that are close to the final output layer learn higher level features, such as the type of the input image. Suppose we have a trained model that works well in one setting, with probability distribution $f_{XY}^{(1)}$, and now we would like to train another model in a different setting, with with probability distribution $f_{XY}^{(2)}$. If the images drawn from the distributions $f_{XY}^{(1)}$ and $f_{XY}^{(2)}$ share some features, then it is possible to transfer weights from the model trained on images sampled from $f_{XY}^{(1)}$, to the model that we would like to train, using images sampled from $f_{XY}^{(2)}$, with the assumption that the weights from the model trained on images sampled from $f_{XY}^{(1)}$, can also be useful in extracting features from images drawn from the distribution $f_{XY}^{(2)}$. So, instead of training the second model from scratch, we can initialize the weights of the

second model to those of the first model that we trained in a different setting (e.g. distribution $f_{XY}^{(1)}$), and utilize the common features we have already learned in the previous setting.

APPENDIX C: BATCH NORMALIZATION

The weights of each layer in a NN model change throughout the training phase, which implies that the activations of each layer will also change. Given that the activations of any given layer are the inputs to the subsequent layer, this means that the input distribution changes at every step. This is far from ideal because it forces each intermediate layer to continuously adapt to its changing inputs. Batch normalization is used to ameliorate this problem by normalizing the activations of each layer. In practice, this is accomplished by adding two trainable parameters to each layer, so the normalized output is multiplied by a standard deviation parameter, and then shifted by a mean parameter. With this approach, only two parameters are changed for each activation, as opposed to losing the stability of the network by changing all the weights. It is expected that through this method each layer will learn on a more stable distribution of inputs, which may accelerate the training stage.

[1]*NCSA, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[2]*Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[3]*Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[4]*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*
[5]*Caltech/IPAC, California Institute of Technology, Pasadena, CA 91125, USA*
[6]*Department of Physics and Astronomy, University of California, Riverside, CA 92507, USA*
[7]*Department of Physics and Astronomy, University of Toledo, Toledo, OH 43606, USA*
[8]*Department of Physics and Astronomy, University of Wyoming, Laramie, WY 82071, USA*
[9]*Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218, USA*
[10]*Centro de Astronomía, Universidad de Antofagasta, Avenida Angamos 601, Antofagasta 1270300, Chile*
[11]*Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Grabengasse 1, D-69117 Heidelberg, Germany*
[12]*Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, D-85748 Garching, Germany*
[13]*Observatories of the Carnegie Institution for Science, Pasadena, CA 91101, USA*
[14]*Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile*
[15]*Las Campanas Observatory, Colina El Pino, Casilla 601, La Serena, Chile*

This paper has been typeset from a TEX/LATEX file prepared by the author.